ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/310658007

Causality analysis detects the regulatory role of maternal effect genes in the early Drosophila embryo

Article *in* Genomics Data · November 2016

DOI: 10.1016/j.gdata.2016.11.013

CITATIONS
0

READS
30

3 authors:

Zara Ghodsi
Bournemouth University
12 PUBLICATIONS **19** CITATIONS
SEE PROFILE

Xu Huang
De Montfort University
8 PUBLICATIONS **6** CITATIONS
SEE PROFILE

Hossein Hassani
94 PUBLICATIONS **1,230** CITATIONS
SEE PROFILE

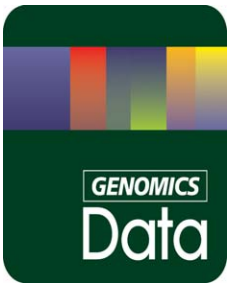
Some of the authors of this publication are also working on these related projects:

developing theoretical aspects of SSA View project

All content following this page was uploaded by Xu Huang on 23 November 2016.

The user has requested enhancement of the downloaded file.

Accepted Manuscript



Causality analysis detects the regulatory role of maternal effect genes in the early *Drosophila* embryo

Zara Ghodsi, Xu Huang, Hossein Hassani

PII: S2213-5960(16)30173-8
DOI: doi:10.1016/j.gdata.2016.11.013
Reference: GDATA 605

To appear in: *Genomics Data*

Received date: 1 May 2016
Revised date: 28 October 2016
Accepted date: 10 November 2016

Please cite this article as: Zara Ghodsi, Xu Huang, Hossein Hassani, Causality analysis detects the regulatory role of maternal effect genes in the early *Drosophila* embryo, *Genomics Data* (2016), doi:10.1016/j.gdata.2016.11.013

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Causality analysis detects the regulatory role of maternal effect genes in the early *Drosophila* embryo

Zara Ghodsi[1,2], Xu Huang[1] and Hossein Hassani [3]

[1] *Statistical Research Centre, Bournemouth University, 89 Holdenhurst Road, Bournemouth BH8 8EB, UK*

[2] *Translational Genetics Group, Bournemouth University, Fern Barrow, Poole, BH125BB, UK*

[3] *Institute for International Energy Studies (IIES), Tehran 1967743 711, Iran*

Abstract

In developmental studies, inferring regulatory interactions of segmentation genetic network plays a vital role in unveiling the mechanism of pattern formation. As such, there exists an opportune demand for theoretical developments and new mathematical models which can result in a more accurate illustration of this genetic network. Accordingly, this paper seeks to extract the meaningful regulatory role of the maternal effect genes using a variety of causality detection techniques and to explore whether these methods can suggest a new analytical view to the gene regulatory networks. We evaluate the use of three different powerful and widely-used models representing time and frequency domain granger causality and convergent cross mapping technique with the results being thoroughly evaluated for statistical significance. Our findings show that the regulatory role of maternal effect genes is detectable in different time classes and thereby the method is applicable to infer the possible regulatory interactions present among the other genes of this network.

Keywords: bicoid; caudal, Drosophila melanogaster, segmentation, time and frequency domain causality, convergent cross mapping.

1 Introduction

Segmentation in *Drosophila melanogaster* is a particularly well studied process which highlights the role of gene regulatory networks (GRNs) in the earliest stage of development [1]. In segmentation GRN, there are three fundamental types of genes which play a crucial role in *Drosophila* development: maternal effect genes, gap genes and pair rule genes [2]. Among them, the maternal effect genes including bicoid (*bcd*) [1] and caudal (*cad*) must be addressed as the most important factors since

[1] In what follows, the italic lower-case *bcd* represents either the gene or mRNA and Bcd refers to protein. This can be applied for all other genes mentioned in this paper (for example, *cad* and Cad)

ACCEPTED MANUSCRIPT

they respectively determine most aspects of anterior and posterior axis of an adult fruit fly and more importantly, they commence the sequential activation of segmentation GRN [2–4].

The segmentation GRN is perhaps the best-studied transcriptional network in *Drosophila* development. Therefore, there are considerable attempts to portrait a picture of the interactions presented between regulators in this GRN. Quantitatively, it is common to model GRNs using ordinary differential equations (ODEs) or stochastic ODEs [5, 6]. Even though, the substantial progress which has been made in modeling transcriptional regulations using these models in recent years is not deniable, the enormous number of regulatory functions obtained by these models and the estimation of parameters which are difficult to assess experimentally can still be considered as two major drawbacks of these methods [7, 8]. Recently, the availability of more data on molecular mechanisms of regulatory interactions has made it possible to study these interactions in more quantitative depth. however, to the best of our knowledge, there is not a particular study which evaluates the dynamic interactions of this system from a statistical causality point of view [9–11]. Hence, this paper seeks to consider an alternative approach based on various causality detection methods to evaluate the possibility of ratifying the validity and reliability of genetic inferences derived from experimental evidences by using proper analytical tools. It is of note that the detected regulatory link can be either inductive (i.e. increasing the protein concentration of one gene raises the protein concentration of the other gene), or inhibitory (i.e. increasing the protein concentration of one gene decreases the protein concentration of the other gene). Any efforts at identifying the nature of the detected interaction would require more extensive research and that objective is beyond the mandate of this paper [12].

The analytical methods used in this paper consist of time and frequency domain granger causality detection (GC) [13] approaches and an advanced non-parametric method - Convergent Cross Mapping (CCM) [14]. Time domain causality test [15] and its developed versions are the most common and generally accepted methods in causal inference analysis. Frequency domain causality test is the extension of time domain causality test on identifying causality for each individual frequency component instead of computing a single measure for the entire causal association. CCM is an advanced non-parametric method that is designed for a dynamical system involving complex interactions. The fundamental concept of CCM is that the information of the driver variable can be recovered from the predator variable, but not vice versa.

It is imperative to note that since providing robust genetic evidence is an important step in reporting genetic causality, among all the interactions between regulators in segmentation GRN, we have narrowed down this study to the interactions between *bcd* and *cad*, *bcd* and Kruppel (*kr*) and *cad* and *kr* genes which their interactions have been previously accredited via laboratory experimental evidences. Accordingly, extracting these links using mentioned causality detection techniques will give us the credit to step further and apply these methods to find the unknown regulatory links between other genes.

The regulatory role of *bcd* has been unveiled by several studies [3, 16]. According to [17] Bcd is one of few proteins which binds both RNA and DNA targets and can be involved in both transcriptional

ACCEPTED MANUSCRIPT

and post transcriptional regulation. Bcd enhances the transcription of anterior gap genes such as *kr* and represses the translation of *cad* in the anterior region of the embryo [16]. In 2002, through an experimental approach, Niessing et al. showed that the translational repression of *cad* mRNA by Bcd depends on a functional eIF4E-binding motif [18]. The *cad* and *kr* genes are also required for a normal segmentation of the embryo. As noted in [19], the interaction of cad and kr gene is an important input of the segmentation genetic network.

In applying causality detection techniques, it should also be noted that as it has been previously shown by several studies, these methods are sensitive to noise [20–22] and gene expression profiles are exceedingly noisy [23]. As it has been shown in Figure 1, the profile achieved by flourocence antibodies technique is highly volatile and in such cases, establishing a cause-and-effect relationship is more challenging and demands applying a noise filtering step prior to causation studies. In order to overcome these issues, among several noise filtering techniques, we have applied Singular Spetrum Analysis (SSA) which is a powerful method and has recently transformed itself into a valuable tool for gene expression signal extraction (see, for example, [24–27]).

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

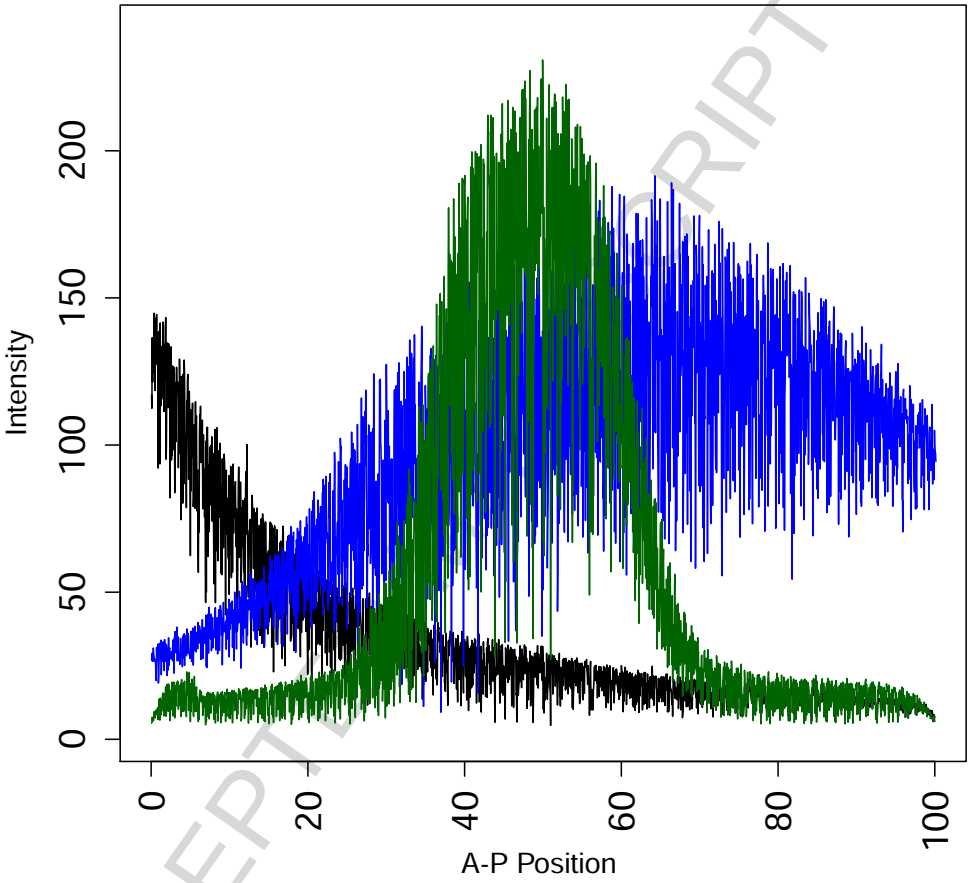


Figure 1: A typical example of noisy Bcd, Cad and Kr for embryo *ms26* at time class 14(1). Black, blue and green colours depict Bcd, Cad and Kr profiles respectively. The x-axis shows the position of the nuclei along the Anterior-Posterior (A-P) axis of the embryo and Y-axis shows the fluorescence intensity levels.

The remainder of this paper is organised such that Section 2 describes the analytical methods used in this study which is followed by description of the data in Section 3. Section 4 summarises the empirical results and the paper concludes with a concise summary in Section 5.

2 Causality Detection and Noise Filtering Techniques

2.1 Time domain Granger causality

Granger causality test [15] is the most generally accepted and significant method for causality analyses in various disciplines. Various applications and developments of this technique, also more specifically

ACCEPTED MANUSCRIPT

in the biomedical area, can be found in [28–36]. The regression formulation of Granger causality states that vector X_i is the cause of vector Y_i if the past values of X_i are helpful in predicting the future value of Y_i, two regressions are considered as follows:

$$Y_i = \sum_{t=1}^{T} \alpha_t Y_{i-t} + \varepsilon_{1i}, \tag{1}$$

$$Y_i = \sum_{t=1}^{T} \alpha_t Y_{i-t} + \sum_{t=1}^{T} \beta_t X_{i-t} + \varepsilon_{2i}, \tag{2}$$

where $i = 1, 2, \cdots, N$ (N is the number of observations), T is the maximal time lag, α and β are vectors of coefficients, ε is the error term. The first regression is the model that predicts Y_i by using the history of Y_i only, while the second regression represents the model of Y_i is predicted by the past information of both X_i and Y_i. Therefore, the conclusion of existing causality is conducted if the second model is a significantly better model than the first one.

2.2 Frequency domain causality

The frequency domain causality test is the extension of time domain GC test that identifys the causality between different variables for each frequency. In order to briefly introduce the testing methodology, we mainly follow [13, 37]. More details can be found in [38].

It is assumed that two dimensional vector containing X_i and Y_i (where $i = 1, 2, \cdots, N$ and N is the number of observations) with a finite-order Vector Auto-regression Model (VAR) representative of order p,

$$\Theta(R) \begin{pmatrix} Y_i \\ X_i \end{pmatrix} = \begin{pmatrix} \Theta_{11}(R) & \Theta_{12}(R) \\ \Theta_{21}(R) & \Theta_{22}(R) \end{pmatrix} \begin{pmatrix} Y_i \\ X_i \end{pmatrix} + \mathcal{E}_i, \tag{3}$$

where $\Theta(R) = I - \Theta_1 R - ... - \Theta_p R_p$ is a 2×2 lag polynomial and $\Theta_1, ..., \Theta_p$ are 2×2 autoregressive parameter matrices, with $R^k X_i = X_{i-k}$ and $R^k Y_i = Y_{i-k}$. The error vector $\mathcal{E}$ is white noise with zero mean, and $E(\mathcal{E}_i \mathcal{E}_i') = \mathbf{Z}$, where $\mathbf{Z}$ is positive definite matrix. The moving average (MA) representative of the system is

$$\begin{pmatrix} Y_i \\ X_i \end{pmatrix} = \Psi(R)\eta_i = \begin{pmatrix} \Psi_{11}(R) & \Psi_{12}(R) \\ \Psi_{21}(R) & \Psi_{22}(R) \end{pmatrix} \begin{pmatrix} \eta_{1i} \\ \eta_{2i} \end{pmatrix}, \tag{4}$$

with $\Psi(R) = \Theta(R)^{-1}\mathbf{G}^{-1}$ and $\mathbf{G}$ is the lower triangular matrix of the Cholesky decomposition $\mathbf{G}'\mathbf{G} = \mathbf{Z}^{-1}$, such that $E(\eta_t \eta_t') = I$ and $\eta_i = \mathbf{G}\mathcal{E}_i$. The causality test developed in [13] can be written as:

$$C_{X \Rightarrow Y}(\gamma) = log\left[1 + \frac{|\Psi_{12}(e^{-i\gamma})|^2}{|\Psi_{11}(e^{-i\gamma})|^2}\right]. \tag{5}$$

However, according to this framework, no Granger causality from X_i to Y_i at frequency γ corresponds to the condition $|\Psi_{12}(e^{-i\gamma})| = 0$, this condition leads to

$$|\Theta_{12}(e^{-i\gamma})| = |\Sigma_{k=1}^{p} \Theta_{k,12} \cos(k\gamma) - i\Sigma_{k=1}^{p} \Theta_{k,12} \sin(k\gamma)| = 0, \tag{6}$$

ACCEPTED MANUSCRIPT

where $\Theta_{k,1,2}$ is the $(1,2)th$ element of Θ_k, such that a sufficient set of conditions for no causality is given by [38]

$$\begin{matrix} \Sigma_{k=1}^{p} \Theta_{k,1,2} \cos(k\gamma) = 0 \\ \Sigma_{k=1}^{p} \Theta_{k,1,2} \sin(k\gamma) = 0 \end{matrix}^{,} \tag{7}$$

Hence, the null hypothesis of no Granger causality at frequency γ can be tested by using a standard F-test for the linear restrictions (7), which follows an $F(2, B-2p)$ distribution, for every γ between 0 and π, with B begin the number of observations in the series.

2.3 Convergent Cross Mapping (CCM)

Convergent Cross Mapping (CCM) is firstly introduced in [14] that aimed at detecting the causation among time series and provide a better understanding of the dynamical systems that have not been covered by other well established methods like Granger causality. CCM has proven to be an advance non-parametric technique for distinguishing causations in a dynamic system that contains complex interactions in biological studies and ecosystems, more details can be found in [14, 39–41]. CCM is briefly introduced below by mainly following [14].

Assume there are two variables X_i and Y_i, for which X_i has a causal effect on Y_i. CCM test will test the causation by evaluating whether the historical record of Y_i can be used to get reliable estimates of X_i. Given a library set of n points (not necessarily to be the total number of observations N of two variables) and here set $i = 1, 2, \cdots, n$, the lagged coordinates are adopted to generate an E-dimensional embedding state space [42, 43], in which the points are the library vector X_i and prediction vector Y_i

$$X_i \quad : \quad \{x_i, x_{i-1}, x_{i-2}, \cdots, x_{i-(E-1)}\}, \tag{8}$$

$$Y_i \quad : \quad \{y_i, y_{i-1}, y_{i-2}, \cdots, y_{i-(E-1)}\}, \tag{9}$$

The $E+1$ neighbors of Y_i from the library set X_i will be selected, which actually form the smallest simplex that contains Y_i as an interior point. Accordingly, the forecast is then conducted by this process, which is the nearest-neighbour forecasting algorithm of simplex projection [43]. The optimal E will be evaluated and selected based on the forward performances of these nearby points in an embedding state space.

Therefore, by adopting the essential concept of Empirical Dynamic Modeling (EDM) and generalized Takens' Theorem [42], two manifolds are conducted based on the lagged coordinates of the two variables under evaluation, which are the attractor manifold M_Y constructed by Y_i and respectively, the manifold M_X by X_i. The causation will then be identified accordingly if the nearby points on M_Y can be employed for reconstructing observed X_i. Note that the correlation coefficient ρ is used for the estimates of cross map skill due to its widely acceptance and understanding, additionally, leave-

ACCEPTED MANUSCRIPT

one-out cross-validation is considered a more conservative method and adopted for all evaluations in CCM.

2.4 Singular Spectrum Analysis

SSA is a powerful non parametric method and has been previously applied for signal extraction of gene expression profiles [24–27].The basic SSA method consists of two complementary stages: decomposition and reconstruction [44]. Throughout the first stage, the gene expression profile is decomposed allowing to differentiate between signal and noise. Throughout the second stage, the less noisy series is reconstructed [45]. A short description of the SSA technique is given below, for more detailed information, see for example, [44, 46].

Step 1: Embedding. Here, the one-dimensional time series $Y_N = (y_1, \ldots, y_N)$ is transferred into the multi-dimensional series $X_1, \ldots, X_K$ with vectors $X_i = (y_i, \ldots, y_{i+L-1})^T \in \mathbf{R}^L$, where $L(2 \leq L \leq N-1)$ is the window length and $K = N - L + 1$. The result of this step is the trajectory matrix $\mathbf{X} = [X_1, \ldots, X_K] = (x_{ij})_{i,j=1}^{L,K}$.

Step 2: SVD. Here, we perform the SVD of $\mathbf{X}$. Denote by $\lambda_1, \ldots, \lambda_L$ the eigenvalues of $\mathbf{X}\mathbf{X}^T$ arranged in the decreasing order ($\lambda_1 \geq \ldots \geq \lambda_L \geq 0$) and by $U_1 \ldots U_L$ the corresponding eigenvectors. The SVD of $\mathbf{X}$ can be written as $\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_L$, where $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^T$.

Step 3: Grouping. The grouping consists in splitting the elementary matrices into several groups and summing the matrices within each group.

Step 4: Diagonal averaging. The purpose of diagonal averaging is to transform a matrix to the form of a Hankel matrix, which can be subsequently converted to a time series.

3 Data

The quantitative *bcd*, *cad* and *kr* gene expression profiles representing the protein concentrations of these genes in wild-type *Drosophila* embryos are achieved using the confocal scanning microscopy of fixed embryos immunostained for segmentation proteins and is available via FlyEx database (http://urchin.spbcas.ru/flyex/). The applied antibody allows the visualisation of the proteins under study. Such quantification relies on the assumption that the actual protein concentrations detected by the antibodies and the fluorescence intensities are linearly related to the embryos natural protein concentration [47, 48].

To this aim, a 1024×1024 pixel confocal image with 8 bits of fluorescence data was obtained for each embryo which then transformed into an ASCII table. The ASCII table contains the fluorescence intensity levels attributed to each nucleus in the %10 of longitudinal strips (i.e. only the nuclei correspondents to the central 10% strip consists of the 45-55% of the dorsoventral (DV) axis are selected) along the A-P axis and is unprocessed for any noise reduction methods. Figure 2 shows an example of a confocal image with the %10 longitudinal strip.

ACCEPTED MANUSCRIPT

Since the segment determination starts from cleavage cycle 10 and lasts until the end of cleavage cycle 14A (when proteins synthesised from maternal transcripts begin to appear up to the onset of gastrulation) the data has been categorised to five main cycles of 10 to 14A. Additionally, as the cleavage cycle 14A is considerably longer in time, to facilitate the analysis, temporal classes 1 to 8 have been considered as the subgroups of this cleavage cycle [47, 48]. It should also be noted that each class of data contains a different number of embryos.

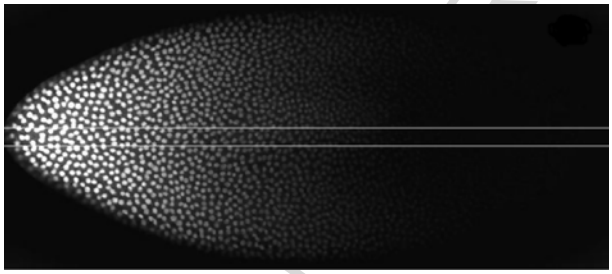


Figure 2: Confocal image of an embryo at time class 14(1). White horizontal lines depict the 10% strip utilised to collect data. Figure adapted from [49].

Table 1 presents the number of embryos studied per each time class. It is of note the expression profile of each embryo has a different length of data where the third column in this table reports the average.

Time class	N	Length	SD
10	5	127	18.83
11	12	276	25.83
12	15	489	97.18
13	47	1224	78.56
14(1)	28	2318	143.87
14(2)	15	2315	86.83
14(3)	20	2367	141.05
14(4)	17	2309	119.16
14(5)	14	2301	126.96
14(6)	18	2347	103.74
14(7)	13	2007	229.61
14(8)	12	1600	311.21

Table 1: Different time classes and the embryos studied per each time class.

Note: N= Number of embryos studied per each time class, Length= The average length of data of expression profiles, SD= Standard deviation of length of data.

Although confocal scanning microscopy is a generally employed technique for measuring the gene expression profiles, its use in systems biology studies presents a number of challenges such as the considerable amount of noise entering data after quantifying the fluorescence intensity. Possible errors

ACCEPTED MANUSCRIPT

in instrument functionality, sample preparation and mathematical treatment of data have been considered as the most common sources of noise [50]. In order to improve the mathematical treatment of data cleaning stage and extracting the signal from the original noisy data, we have applied SSA. Figure 3 illustrates the output from this effort. It is evident that the SSA method provides a relatively smooth signal line with correlation below 0.10 which credits the satisfactory level of separation between noise and signal using SSA [27].

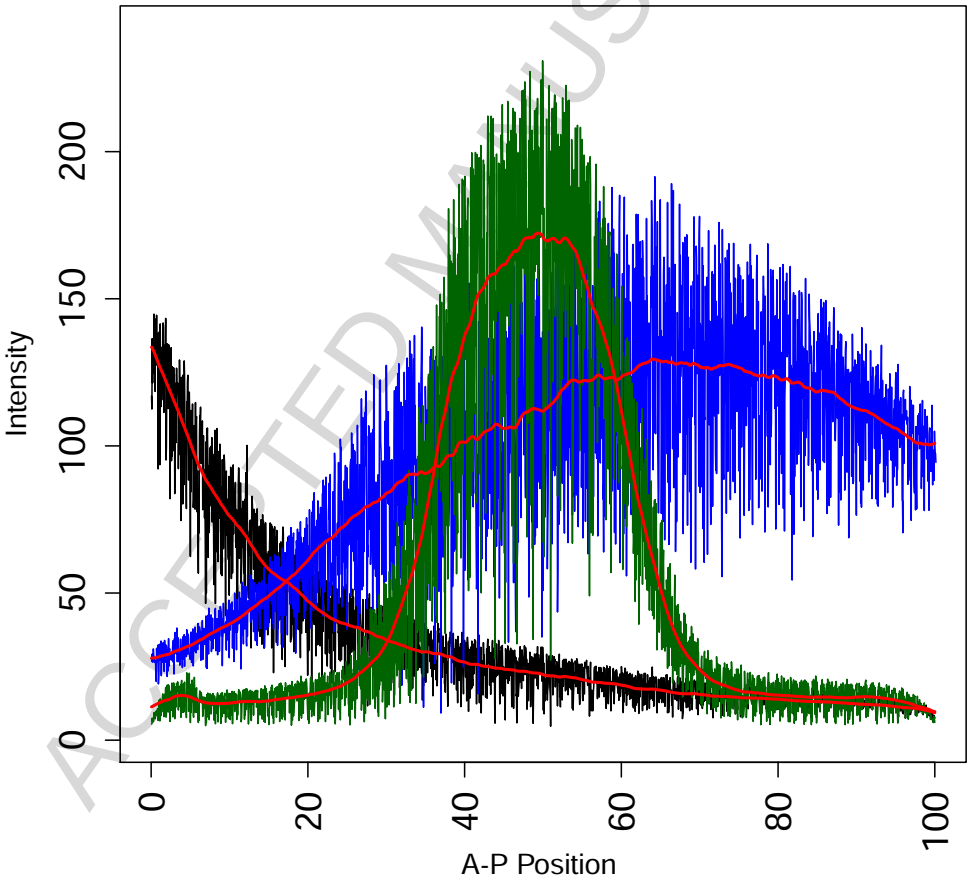


Figure 3: A typical example of noisy Bcd, Cad and Kr along with the extracted signals in red for embryo *ms26* at time class 14(1). Black, blue and green colours depict Bcd, Cad and Kr profiles respectively. The x-axis shows the position of the nuclei along the Anterior-Posterior (A-P) axis of the embryo and Y-axis shows the fluorescence intensity level.

ACCEPTED MANUSCRIPT

4 Empirical Results

This section provides a summary of the results following applying the three causality detection approaches before and after filtering the expression profiles using SSA. For all evaluations, we have ensured that all the test requirements are satisfied by choosing the optimal indices. Table 2 illustrates the findings of the causality detection analysis on Bcd and Cad profiles, where"Yes" stands for the detected regulatory relationship by the adopted test. The p-values reported for time domain GC test are the average p-values attained for each time class. For time domain GC test, the co-integration test is conducted only for those variables having one unit root. Since none of the tested groups showed significant results in indicating co-integration, the co-integration test result is not reported here. The optimal lag for each VAR model is selected by comparing the information criteria matrix, which includes results based on the AIC [51], HQ [52], SIC [53] and FPE [54] criteria.

Table 2: A summary of the causality tests results for Bcd on Cad profiles.

Time Class	**Time Domain GC**				**Frequency Domain GC**		**CCM**	
	Noisy Series		Filtered Series		Noisy Series	Filtered Series	Noisy Series	Filtered Series
	YES/NO	p-value	YES/NO	p-value	YES/NO	YES/NO	YES/NO	YES/NO
10	NO	0.68	NO	0.45	NO	YES	YES	YES
11	NO	0.71	NO	0.33	NO	YES	YES	YES
12	NO	0.89	NO	0.32	NO	YES	YES	YES
13	NO	0.89	NO	0.24	NO	YES	YES	YES
14(1)	NO	0.95	YES	0.05	NO	YES	YES	YES
14(2)	NO	0.98	YES	0.04	NO	YES	YES	YES
14(3)	NO	0.98	YES	0.01	NO	YES	YES	YES
14(4)	NO	0.94	YES	0.01	NO	YES	YES	YES
14(5)	NO	0.95	YES	0.00	NO	YES	YES	YES
14(6)	NO	0.96	YES	0.00	NO	YES	YES	YES
14(7)	NO	0.81	YES	0.00	NO	YES	YES	YES
14(8)	NO	0.79	YES	0.04	NO	YES	YES	YES

Note: Differentiations are taken accordingly for stationarity prior to the tests;
Optimal lag lengthes are chosen based on the AIC, HQ, SIC and FPE criterions. "Yes" stands for the detected regulatory link and "No" means the regulatory link could not be detected by the adopted test.

According to Table 2, it is evident that there is a significant difference in results before and after reducing the noise from the profiles. The regulatory link between Bcd and Cad can be detected by neither time domain nor frequency domain tests in the presence of noise. Accordingly, it is clear that the filtering capability displayed by SSA is indeed advantageous for causality detection analysis.

Nevertheless, as can be seen, the feasibility of capturing the regulatory link for CCM method has not been affected by noise and the results achieved by this test confirm the regulatory relationship between Bcd and Cad in expression profiles with and without noise. However, regardless of the time class, the index representing the ability of cross mapping is relatively smaller on average for noisy series than filtered series.

ACCEPTED MANUSCRIPT

It is of note that the length of the data under study vary between different time classes. Time class 10 to 13 and 14(7-8) have shorter lengths comparing to the time classes 14(1-6), which may be the reason of getting slightly smaller p-values for time class 11 to 13 and 14(8) comparing to the rest of the sub classes of time class 14. Yet, the frequency domain test shows less sensitivity to the data length possibly because this method identifies the possible regulative link for each individual frequency component rather than the entire series.

Furthermore, the p-values obtained for both noisy and filtered data of all the embryos in different time classes are summarised in Figures 4 and 5 as box and whisker diagram respectively. They follow the standard format of box plot on displaying the distribution of the p-values based on maximum, upper quartile, median, lower quartile, and minimum. A close look at Figures 4 and 5 suggests that the time domain GC test cannot detect any regulatory link in the presence of noise, while the results for filtered series are significant and more consistent especially for those time classes after 14(1). Comparing the p-values illustrated in Figure 4 and 5, it is evident that the length of the series and level of intensities have more effect on the result of the noisy data than the filtered one as the p-values in Figure 4 are getting more insignificant for the final subclasses of time class 14, where there is a decreasing pattern for these two parameters in the expression profiles. Likewise, for the frequency domain GC test, the links have been detected for all the filtered series, whilst there is no regulatory relationship detected for non-filtered ones.

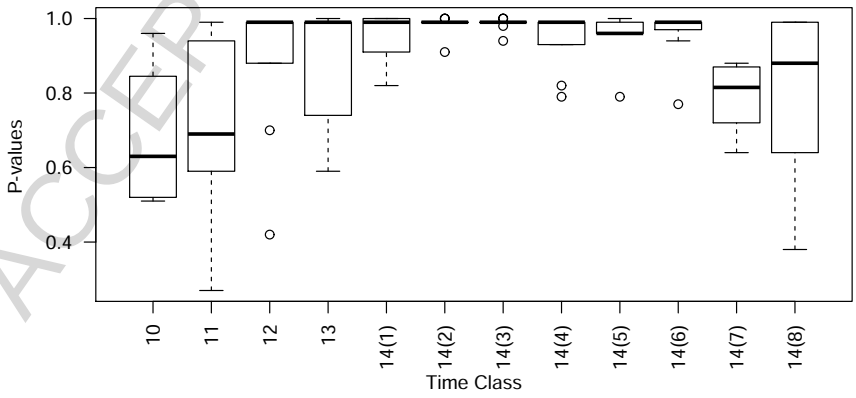


Figure 4: Box Plots of Time Domain GC Test P-values for Noisy Series. (Circle refers to the corresponding outlier that is more/less than 1.5 times of upper/lower quartile; the central rectangle spans the upper quartile to the lower quartile; the segment inside the rectangle indicates the median; whiskers above and below the box refer to the maximum and minimum.)

ACCEPTED MANUSCRIPT

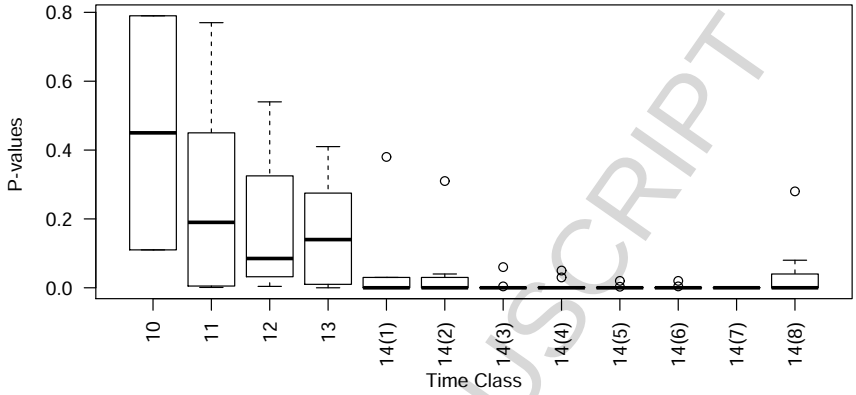


Figure 5: Box Plots of Time Domain GC Test P-values for Filtered Series. (Circle refers to the corresponding outlier that is more/less than 1.5 times of upper/lower quartile; the central rectangle spans the upper quartile to the lower quartile; the segment inside the rectangle indicates the median; whiskers above and below the box refer to the maximum and minimum.)

Tables 3 and 4 present the results of the conducted analysis to detect the regulatory link between Bcd and Kr profiles and Cad and kr profiles respectively. As can be seen, reducing the noise level is an essential step in detecting the regulatory link using the time domain and frequency domain tests. Similar to the results reported in Table 2, CCM method can again efficiently identify the regulatory relationship even in the presence of noise.

Table 3: A summary of the causality tests results for Bcd on Kr profiles.

	Time Domain GC				**Frequency Domain GC**		**CCM**	
Time Class	Noisy Series		Filtered Series		Noisy Series	Filtered Series	Noisy Series	Filtered Series
	YES/NO	p-value	YES/NO	p-value	YES/NO	YES/NO	YES/NO	YES/NO
12	NO	0.71	NO	0.15	NO	YES	YES	YES
13	NO	0.66	YES	0.04	NO	YES	YES	YES
14(1)	NO	0.89	YES	0.03	NO	YES	YES	YES
14(2)	NO	0.93	YES	0.01	NO	YES	YES	YES
14(3)	NO	0.97	YES	0.01	NO	YES	YES	YES
14(4)	NO	0.94	YES	0.00	NO	YES	YES	YES
14(5)	NO	0.95	YES	0.00	NO	YES	YES	YES
14(6)	NO	0.92	YES	0.00	NO	YES	YES	YES
14(7)	NO	0.81	YES	0.00	NO	YES	YES	YES

Note: Differentiations are taken accordingly for stationarity prior to the tests;
Optimal lag lengthes are chosen based on the AIC, HQ, SIC and FPE criterions. "Yes" stands for the detected regulatory link and "No" means the regulatory link could not be detected by the adopted test.

ACCEPTED MANUSCRIPT

Table 4: A summary of the causality tests results for Cad on Kr profiles.

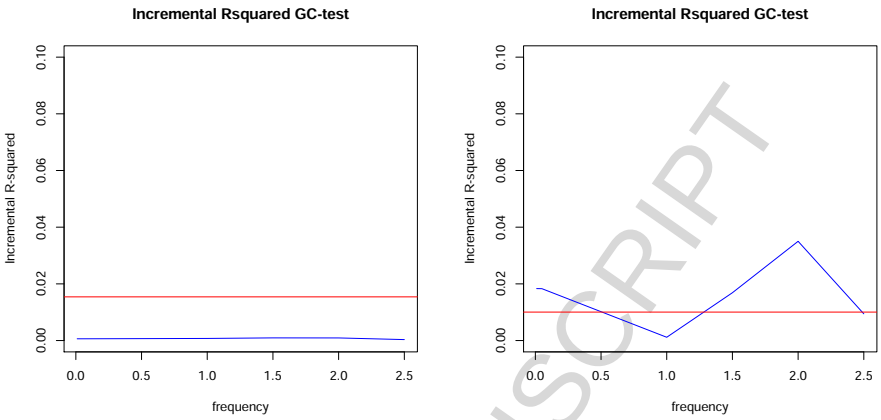
Time Class	Time Domain GC				Frequency Domain GC		CCM	
	Noisy Series		Filtered Series		Noisy Series	Filtered Series	Noisy Series	Filtered Series
	YES/NO	p-value	YES/NO	p-value	YES/NO	YES/NO	YES/NO	YES/NO
12	NO	0.39	NO	0.25	NO	YES	YES	YES
13	NO	0.78	NO	0.11	NO	YES	YES	YES
14(1)	NO	0.84	YES	0.05	NO	YES	YES	YES
14(2)	NO	0.89	YES	0.03	NO	YES	YES	YES
14(3)	NO	0.94	YES	0.01	NO	YES	YES	YES
14(4)	NO	0.91	YES	0.01	NO	YES	YES	YES
14(5)	NO	0.87	YES	0.00	NO	YES	YES	YES
14(6)	NO	0.82	YES	0.00	NO	YES	YES	YES
14(7)	NO	0.75	YES	0.00	NO	YES	YES	YES

Note: Differentiations are taken accordingly for stationarity prior to the tests;
Optimal lag lengthes are chosen based on the AIC, HQ, SIC and FPE criterions. "Yes" stands for the detected regulatory link and "No" means the regulatory link could not be detected by the adopted test.

Figures 6, 7 and 8 depict an example of the results obtained by frequency domain GC test for Bcd–Cad, Bcd–Kr and Cad–Kr profile pairs respectively [2]. In these figures, the blue line represents the statistic test of each specific frequency, and the red line represents the 5% critical value for all the frequencies. The horizontal axis gives the parameter w to calculate the corresponding frequency f by $f = 2\pi/w$. Therefore, when the test statistics is above or very close to the 5% critical value, the causality is detected for that corresponding frequency. As the component of each frequency is considered separately for identifying possible causal link, the impacts of relatively less information are significantly reduced. However, there are some results of filtered series showing very minor differences between the test statistics and the 5% critical value.

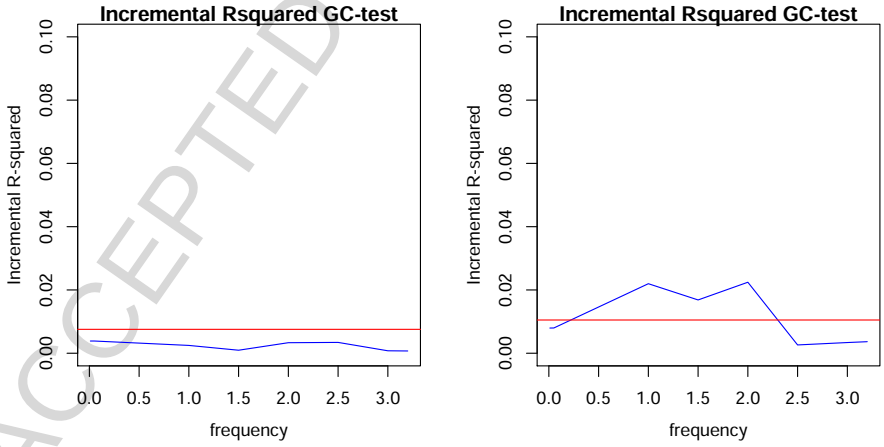
[2]The frequency domain GC test results for all considered pairs of genes related to all different time classes can be found in Appendix 1.

ACCEPTED MANUSCRIPT



Noisy-t11-bcd on cad Filtered-t11-bcd on cad

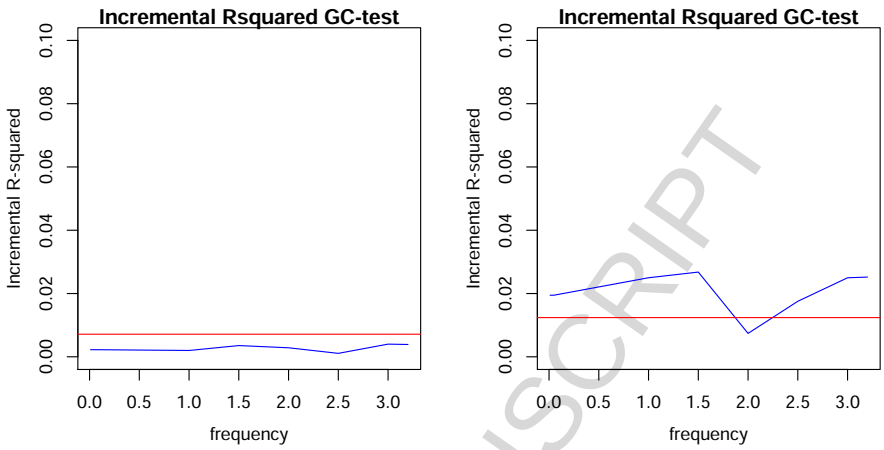
Figure 6: Frequency domain causality test results for Bcd and Cad before and after filtering (time class 11). The blue line represents the statistic test of each specific frequency, and the red line represents the 5% critical value for all the frequencies.



Noisy-t12-bcd on kr Filtered-t12-bcd on kr

Figure 7: Frequency domain causality test results for Bcd and Kr before and after filtering (time class 12).The blue line represents the statistic test of each specific frequency, and the red line represents the 5% critical value for all the frequencies.

ACCEPTED MANUSCRIPT



Noisy-t12-cad on kr Filtered-t12-cad on kr

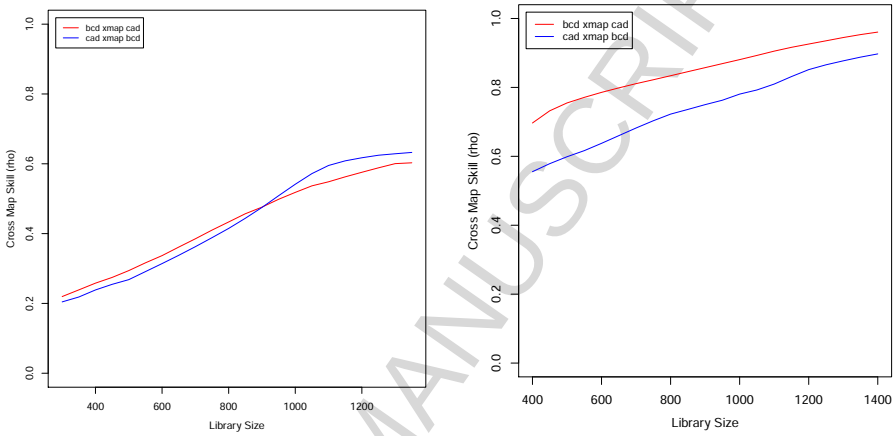
Figure 8: Frequency domain causality test results for Cad and Kr before and after filtering (time class 12).The blue line represents the statistic test of each specific frequency, and the red line represents the 5% critical value for all the frequencies.

For CCM test, the optimal embedding dimension E has been selected for each pair of gene expression profiles based on the nearest neighbor forecasting performance by simplex projection. Figures 9, 10 and 11 represent the examples of the CCM test result for Bcd–Cad, Bcd–Kr and Cad–Kr before and after filtering the profiles [3], where for example regrding the Figure 9, the red line indicates the reconstruction ability of Bcd cross mapping Cad, while the blue line represents the performance of using historical information of Cad on cross mapping Bcd. In general, the higher ability of factor X on reconstructing the attractor reflects more significant causal effects of the attractor on X. The results of CCM reflect close relationships between Bcd and Cad with and without filtering, whilst Bcd shows more significant relationship with Kr comparing to Cad for both original and filtered data. The crossmap abilities of Bcd and Cad on Kr are fairly similar, however, Kr clearly indicates higher reconstruction ability on Bcd comparing to Cad. In more details regarding the relationship between Bcd and Cad, considering the average reconstruction ability represented by ρ, it is suggested that CCM is not affected by the smaller length of the series related to the initial time. However, the increasing pattern of the average level of cross-mapping ability up to time class 14(3), which follows by a decreasing trend for the rest of the subclasses, indicates less accuracy of the results for higher time classes. The approximate average value of ρ over 0.5 for noisy series indicates significant cross-mapping (or reconstruction) ability to identify the causal links. Correspondingly, an average is found to be approximately over 0.8, which reflects stronger causal links detected between Bcd and Cad after filtering. Regarding the relationships between Bcd and Kr, both original and filtered series indicate stronger cross-mapping ability from Kr to Bcd, which means that Bcd shows a more powerful regulatory effect

[3]The CCM test results for all considered pairs of genes related to all different time classes can be found in Appendix 2.

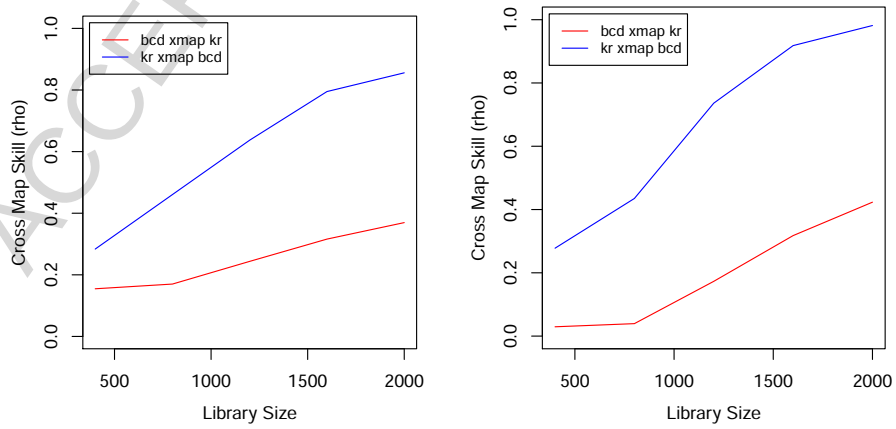
ACCEPTED MANUSCRIPT

on Kr than the other way around. However, this link is slightly more significant in the filtered profiles. In the case of Cad and Kr, the regulatory relationship identified is less significant comparing to the other pairs of genes considered in this study and the average of 0.4 for filtered profiles compared to the average of 0.2 for original series highlights the role of the SSA in improving the achieved results.



Noisy-t14(8)-ccm Filtered-t14(8)-ccm

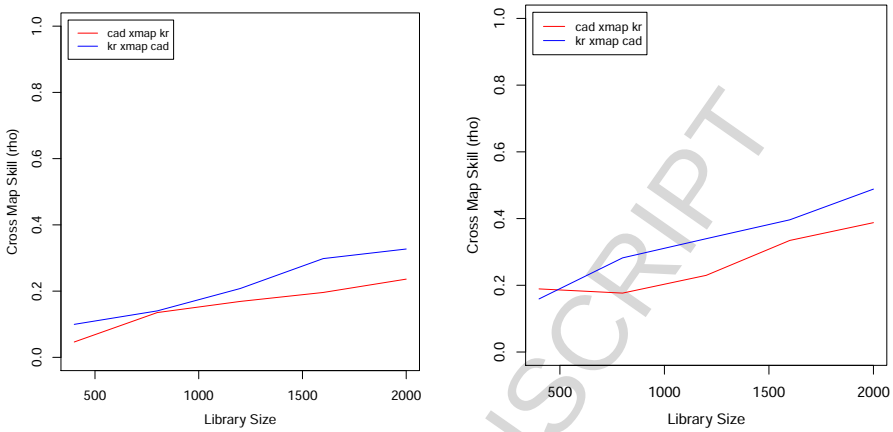
Figure 9: CCM test results for Bcd and Cad before and after filtering (time class 14(8)). The red line indicates the reconstruction ability of Bcd crossmap Cad, while the blue line represents the performance of Cad on crossmapping Bcd.



Noisy-t14(7)-ccm Filtered-t14(7)-ccm

Figure 10: CCM test results for Bcd and Kr before and after filtering (time class 14(7)). The red line indicates the reconstruction ability of Bcd crossmap Kr, while the blue line represents the performance of Kr on crossmapping Bcd.

ACCEPTED MANUSCRIPT



Noisy-t14(5)-ccm Filtered-t14(5)-ccm

Figure 11: CCM test results for Cad and Kr before and after filtering (time class 14(5)). The red line indicates the reconstruction ability of Cad crossmap Kr, while the blue line represents the performance of Kr on crossmapping Cad.

It is of note that the overall findings of this research are consistent with the previous efforts in mathematical modelling the segmentation network [55–57]. For example, [55] presents a succesful canalization study of four gap genes hunchback (*hb*), giant (*gt*) knirps (*kni*) kr using the gene circuit method which uses the concentration of *bcd*, *cad* tailless (*tll*) and genes as outside inputs.

5 Conclusion

Even though the regulatory role of *bcd* on *cad*, *bcd* on *kr* and *cad* on *kr* genes have been previously reported through several genetics experiments, in practice they have not been validated using any causality detection methods. Hence, extracting the regulatory links between these expression profiles were central to this study. We therefore tested various models using the real data to ensure the validity of the findings. We have applied the three causality detection approaches before and after filtering the expression profiles. According to the obtained results the accuracy of data is of critical importance for the success of causality detection studies. Using time domain and frequency domain GC tests, the regulatory link can be detected only after removing the noise from the expression profiles which indicates having an even small amount of error in mean intensities may lead us to obtain a false negative result.

It is also imperative to note that for all pairs of genes considered in this study, the time domain GC fails to detect the regulatory link in time classes 10-13. The poor performance of this model here can be attributed to either the length of the data or low expression level for those time classes. The protein molecules synthesised from maternal transcripts just begin to appear from time class 10 and the number of these morphogens, in the areas where they were concentrated, is at a lower amount for

ACCEPTED MANUSCRIPT

time classes 10-13 comparing to the higher time classes.

According to the achieved results, confirming that there is a regulatory link between *bcd* and *cad*, *bcd* and *kr* and also *cad* and *kr*, it is worth mentioning that the combined application of our filtering method and the causality methods developed in this work provide means to correct errors and hereby makes it possible to obtain more accurate information from expression profiles. This can be easily adapted to the other pairs of genes and is also applicable to a wider range of GRNs to infer the regulatory interactions presented among the genes of that network.

References

[1] Lewis, E. B. (1978). A gene complex controlling segmentation in Drosophila. In Genes, Development and Cancer (pp. 205-217). Springer US.

[2] Bieler, J., Pozzorini, C., & Naef, F. (2011). Whole-embryo modeling of early segmentation in Drosophila identifies robust and fragile expression domains. Biophysical Journal, 101(2), 287-296.

[3] Berleth, T., Burri, M., Thoma, G., Bopp, D., Richstein, S., Frigerio, G., ... & Nsslein-Volhard, C. (1988). The role of localization of bicoid RNA in organizing the anterior pattern of the Drosophila embryo. The EMBO Journal, 7(6), 1749.

[4] Copf, T., Schrder, R., & Averof, M. (2004). Ancestral role of caudal genes in axis elongation and segmentation. Proceedings of the National Academy of Sciences of the United States of America, 101(51), 17711-17715.

[5] Karlebach, G., & Shamir, R. (2008). Modelling and analysis of gene regulatory networks. Nature Reviews Molecular Cell Biology, 9(10), 770-780.

[6] De Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. Journal of Computational Biology, 9(1), 67-103.

[7] Schlitt, T., & Brazma, A. (2007). Current approaches to gene regulatory network modelling. BMC Bioinformatics, 8(Suppl 6), S9.

[8] Wilczynski, B., & Furlong, E. E. (2010). Challenges for modeling global gene regulatory networks during development: insights from Drosophila. Developmental Biology, 340(2), 161-169.

[9] Frigerio, G., Burri, M., Bopp, D., Baumgartner, S., & Noll, M. (1986). Structure of the segmentation gene paired and the Drosophila PRD gene set as part of a gene network. Cell, 47(5), 735-746.

[10] Chaves, M., Albert, R., & Sontag, E. D. (2005). Robustness and fragility of Boolean models for genetic regulatory networks. Journal of Theoretical Biology, 235(3), 431-449.

ACCEPTED MANUSCRIPT

[11] Levine, M., & Davidson, E. H. (2005). Gene regulatory networks for development. Proceedings of the National Academy of Sciences of the United States of America, 102(14), 4936-4942.

[12] Davidson, E., & Levin, M. (2005). Gene regulatory networks. Proceedings of the national academy of sciences of the United States of America, 102(14), 4935-4935.

[13] Geweke, J. (1982). Measurement of linear dependence and feedback between multiple time series. Journal of the American Statistical Association, 77, 304-324.

[14] Sugihara, G., May, R., Ye, H., Hsieh, C. H., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems. Science, 338(6106), 496-500.

[15] Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society, 37(3), 424-438.

[16] Lopes, F. J., Spirov, A. V., & Bisch, P. M. (2012). The role of Bicoid cooperative binding in the patterning of sharp borders in Drosophila melanogaster. Developmental Biology, 370(2), 165-172.

[17] Baird-Titus, J. M., Clark-Baldwin, K., Dave, V., Caperelli, C. A., Ma, J., & Rance, M. (2006). The Solution Structure of the Native K50 Bicoid Homeodomain Bound to the Consensus TAATCC DNA-binding Site. Journal of Molecular Biology, 356(5), 11371151.

[18] Niessing, D., Blanke, S., & Jckle, H. (2002). Bicoid associates with the 5-cap-bound complex of caudal mRNA and represses translation. Genes & Development, 16(19), 2576-2582.

[19] Liu, S., & Jack, J. (1992). Regulatory interactions and role in cell type specification of the Malpighian tubules by the cut, Krppel, and caudal genes of Drosophila. Developmental biology, 150(1), 133-143.

[20] Hiemstra, C., & Jones, J. D. (1994). Testing for linear and nonlinear Granger causality in the stock price-volume relation. The Journal of Finance, 49(5), 1639-1664.

[21] Ancona, N., Marinazzo, D., & Stramaglia, S. (2004). Radial basis function approach to nonlinear Granger causality of time series. Physical Review E, 70(5), 056221.

[22] Zou, C., & Feng, J. (2009). Granger causality vs. dynamic Bayesian network inference: a comparative study. BMC Bioinformatics, 10(1), 1.

[23] Golyandina, N. E., Holloway, D. M., Lopes, F. J., Spirov, A. V., Spirova, E. N., & Usevich, K. D. (2012). Measuring gene expression noise in early Drosophila embryos: nucleus-to-nucleus variability. Procedia Computer Science, 9, 373-382.

[24] Holloway, D. M., Harrison, L. G., Kosman, D., VanarioAlonso, C. E., & Spirov, A. V. (2006). Analysis of pattern precision shows that Drosophila segmentation develops substantial independence from gradients of maternal gene products. Developmental Dynamics, 235(11), 2949-2960.

ACCEPTED MANUSCRIPT

[25] Hassani, H., & Ghodsi, Z. (2014). Pattern recognition of gene expression with singular spectrum analysis. Medical Sciences, 2(3), 127-139.

[26] Ghodsi, Z., Silva, E. S., & Hassani, H. (2015). Bicoid signal extraction with a selection of parametric and nonparametric signal processing techniques. Genomics, Proteomics & Bioinformatics, 13(3), 183-191.

[27] Ghodsi, Z., Hassani, H., & McGhee, K. (2015). Mathematical approaches in studying bicoid gene. Quantitative Biology, 3(4), 182-192.

[28] Sims, C. A. (1972). Money, income, and causality. The American Economic Review, 62(4), 540-552.

[29] Hsiao, C. (1981). Autoregressive modelling and money-income causality detection. Journal of Monetary Economics, 7(1), 85-106.

[30] Sims, C. A., Stock, J. H., & Watson, M. W. (1990). Inference in linear time series models with some unit roots. Econometrica: Journal of the Econometric Society, 58(1), 113-144.

[31] Toda, H. Y., & Phillips, P. C. (1993). Vector autoregressions and causality. Econometrica: Journal of the Econometric Society, 61(6), 1367-1393.

[32] Toda, H. Y., & Yamamoto, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes. Journal of Econometrics, 66(1), 225-250.

[33] Pesaran, H. H., & Shin, Y. (1998). Generalized impulse response analysis in linear multivariate models. Economics Letters, 58(1), 17-29.

[34] Chen, Y., Bressler, S. L., & Ding, M. (2006). Frequency decomposition of conditional Granger causality and application to multivariate neural field potential data. Journal of Neuroscience Methods, 150(2), 228-237.

[35] Gow, D. W., Segawa, J. A., Ahlfors, S. P., & Lin, F. H. (2008). Lexical influences on speech perception: a Granger causality analysis of MEG and EEG source estimates. Neuroimage, 43(3), 614-623.

[36] Deshpande, G., Sathian, K., & Hu, X. (2010). Effect of hemodynamic variability on Granger causality analysis of fMRI. Neuroimage, 52(3), 884-896.

[37] Ciner, C. (2011). Eurocurrency interest rate linkages: A frequency domain analysis. Review of Economics and Finance, 20(4), 498-505.

[38] Breitung, J., & Candelon, B. (2006). Testing for short- and long-run causality: A frequency-domain approach. Journal of Econometrics, 132, 363-378.

ACCEPTED MANUSCRIPT

[39] Deyle, E., Fogarty, M., Hsieh, C., Kaufman, L., MacCall, A., Munch, S., Perretti, C., Ye, H., & Sugihara, G. (2013). Predicting climate effects on Pacific sardine. Proceedings of the National Academy of Sciences, 110(16), 6430-6435.

[40] Ye, H., Deyle, E., Gilarranz, L., & Sugihara, G. (2015). Distinguishing time-delayed causal interactions using convergent cross mapping. Scientific Reports, 5, 14750.

[41] Clark, A. T., Ye, H., Isbell, F., Deyle, E., Cowles, J., Tilman, G., & Sugihara, G. (2015). Spatial convergent cross mapping to detect causal relationships from short time series. Ecology, 96(5), 1174-1181.

[42] Takens, F. (1981). Detecting strange attractors in turbulence Dynamical Systems and Turbulence. Dynamic Systems and Turbulence, 898, 366-381.

[43] Sugihara, G., & May, R. (1990). Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. Nature, 344(6268), 734-741.

[44] Hassani, H. (2007). Singular spectrum analysis: Methodology and comparison. Journal of Data Science, 5(2), 239-257.

[45] Hassani, H., Heravi, H., & Zhigljavsky, A. (2009). Forecasting European industrial production with Singular Spectrum Analysis. International Journal of Forecasting, 25(1), 103-118.

[46] Sanei, S., & Hassani, H. (2015). Singular spectrum analysis of biomedical signals. CRC Press.

[47] Pisarev, A., Poustelnikova, E., Samsonova, M., & Reinitz, J. (2009). FlyEx, the quantitative atlas on segmentation gene expression at cellular resolution. Nucleic Acids Research, 37(suppl 1), D560-D566.

[48] Poustelnikova, E., Pisarev, A., Blagov, M., Samsonova, M., & Reinitz, J. (2004). A database for management of gene expression data in situ. Bioinformatics, 20(14), 2212-2221.

[49] Surkova, S., Kosman, D., Kozlov, K., Myasnikova, E., Samsonova, A.A., Spirov, A., Vanario-Alonso, C.E., Samsonova, M. and Reinitz, J., 2008. Characterization of the Drosophila segment determination morphome. Developmental biology, 313(2), pp.844-862.

[50] Myasnikova, E., Surkova, S., Panok, L., Samsonova, M., & Reinitz, J. (2009). Estimation of errors introduced by confocal imaging into the data on segmentation gene expression in Drosophila. Bioinformatics, 25(3), 346-352.

[51] Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika, 60(2), 255-265.

[52] Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. Journal of the Royal Statistical Society: Series B (Methodological), 42(1), 190-195.

ACCEPTED MANUSCRIPT

[53] Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, 6(2), 461-464.

[54] Akaike, H. (1969). Fitting autoregressive models for prediction. Annals of the institute of Statistical Mathematics, 21(1), 243-247.

[55] Surkova, S., Spirov, A.V., Gursky, V.V., Janssens, H., Kim, A.R., Radulescu, O., Vanario-Alonso, C.E., Sharp, D.H., Samsonova, M. and Reinitz, J., 2009. Canalization of gene expression in the Drosophila blastoderm by gap gene cross regulation. PLoS Biol, 7(3), p.e1000049.

[56] Gursky, V.V., Panok, L., Myasnikova, E.M., Manu, M., Samsonova, M.G., Reinitz, J. and Samsonov, A.M., 2011. Mechanisms of gap gene expression canalization in the Drosophila blastoderm. BMC systems biology, 5(1), p.1.

[57] Surkova, S., Spirov, A.V., Gursky, V.V., Janssens, H., Kim, A.R., Radulescu, O., Vanario-Alonso, C.E., Sharp, D.H., Samsonova, M. and Reinitz, J., 2009. Canalization of gene expression and domain shifts in the Drosophila blastoderm by dynamical attractors. PLoS Comput Biol, 5(3), p.e1000303.

Appendix 1. Frequency Domain GC Test Results

Note that some results of filtered series show a minor difference between the test statistics and the 5% critical value, which is hard to depict in the outcome test plots when considering the same legend for comparison.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

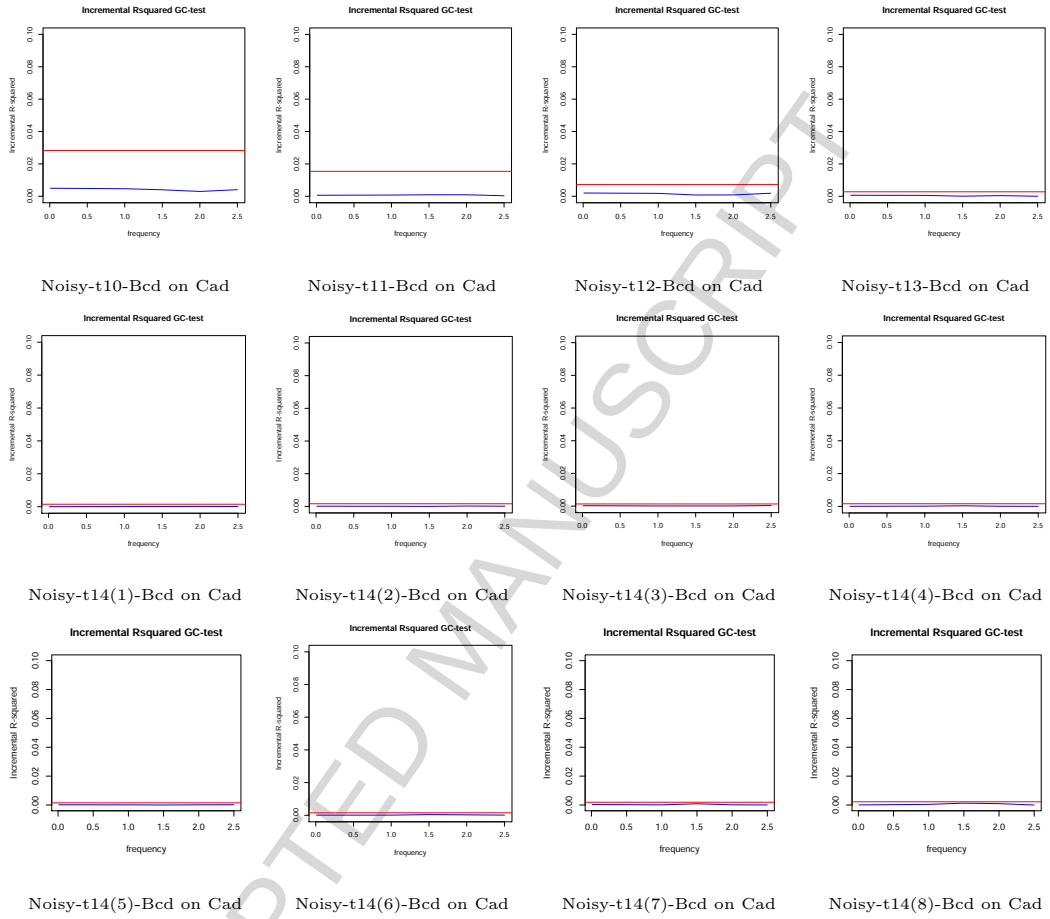


Figure 12: Frequency domain causality test results for Bcd and Cad (Noisy Series).

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

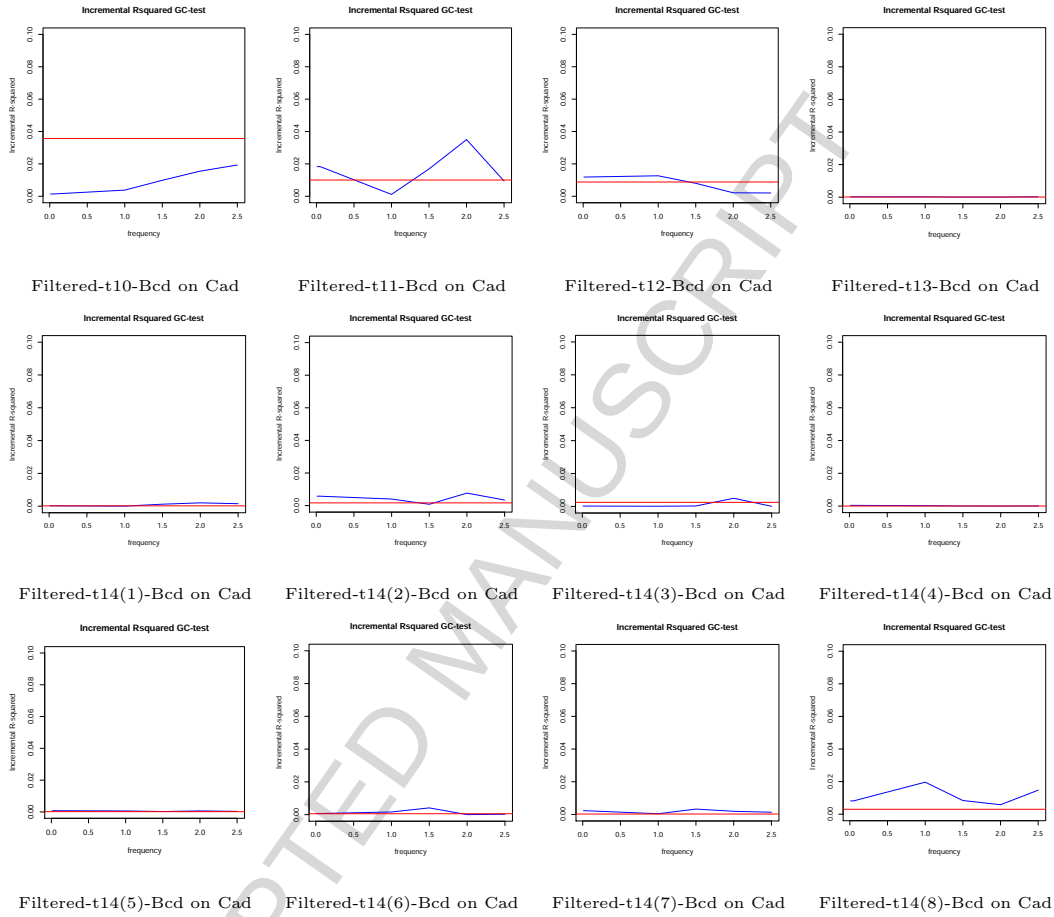


Figure 13: Frequency domain causality test results for Bcd and Cad (Filtered Series).

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

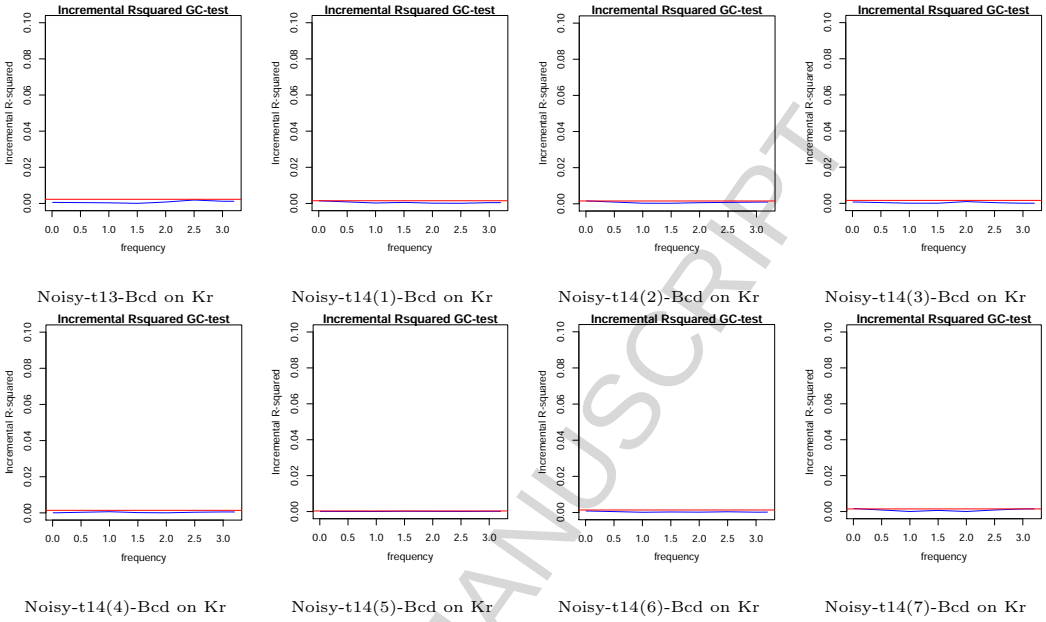


Figure 14: Frequency domain causality test results for Bcd and Kr (Noisy Series).

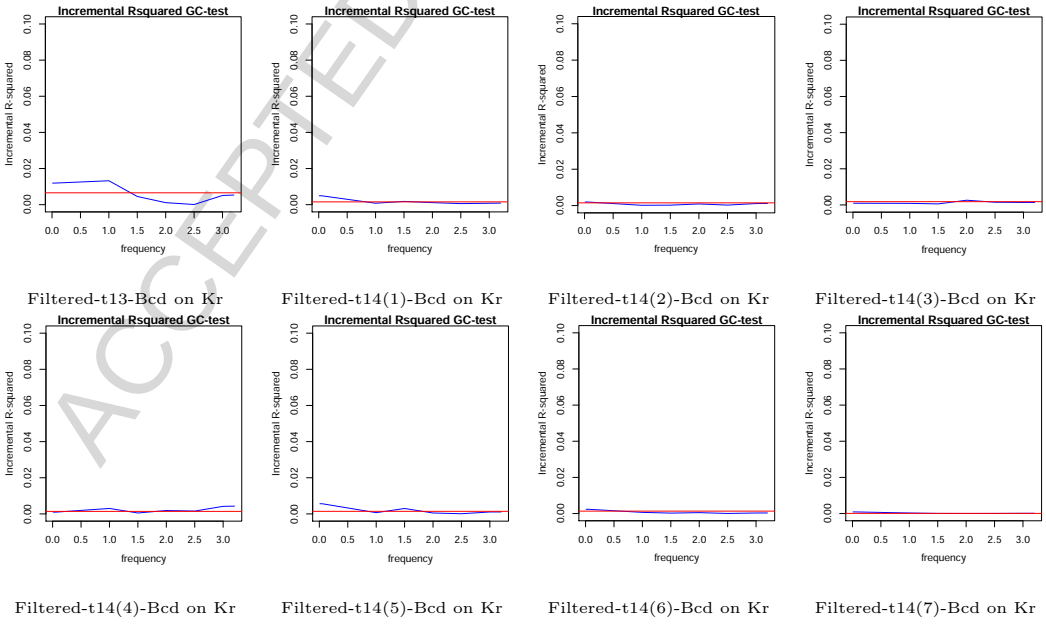


Figure 15: Frequency domain causality test results for Bcd and Kr (Filtered Series).

ACCEPTED MANUSCRIPT

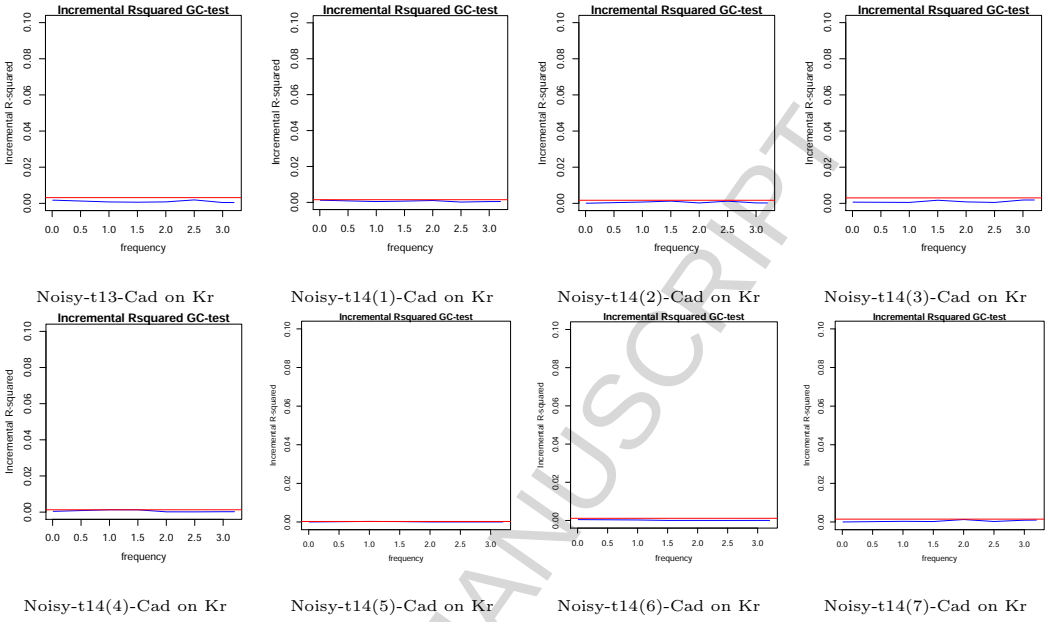


Figure 16: Frequency domain causality test results for Cad and Kr (Noisy Series).

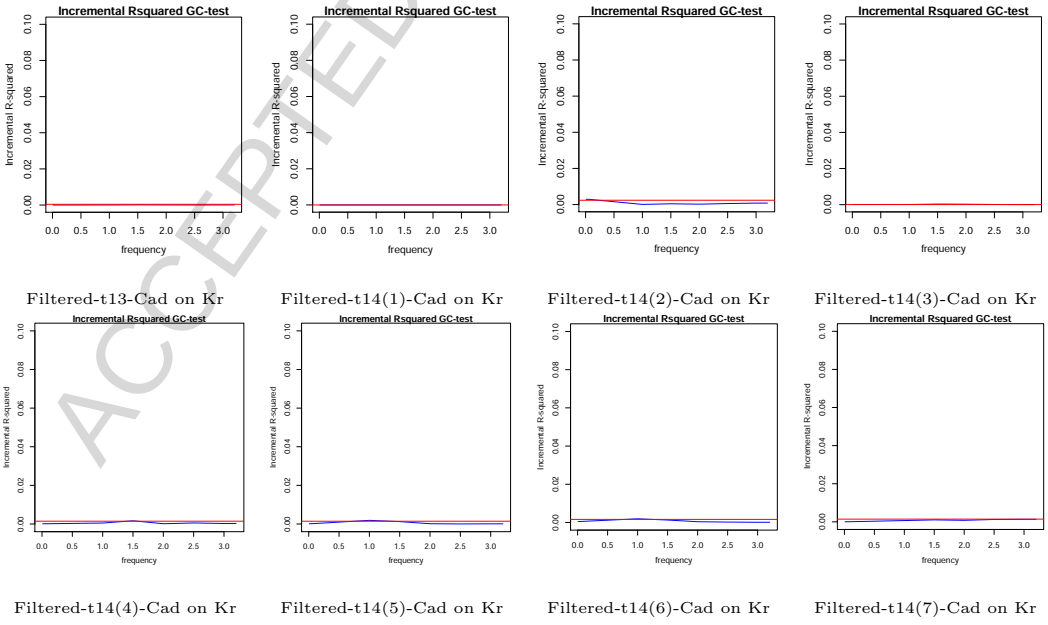
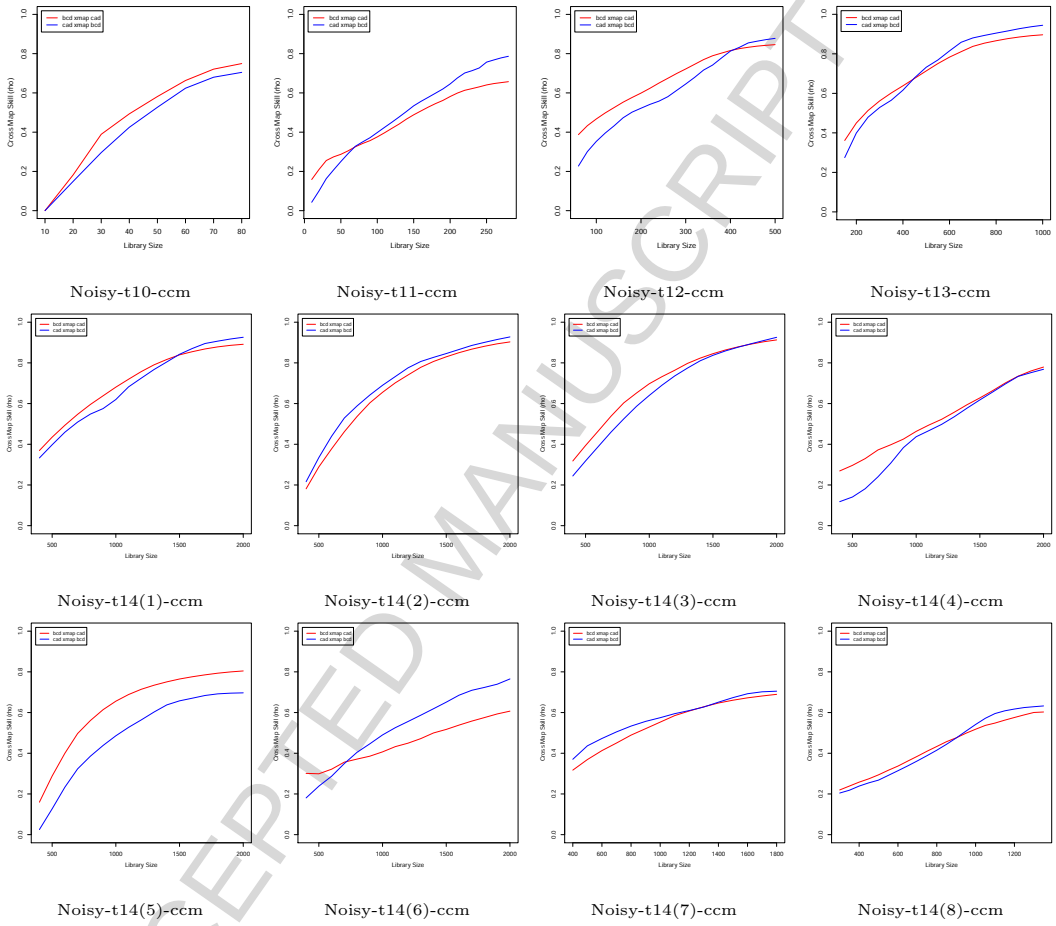


Figure 17: Frequency domain causality test results for Cad and Kr (Filtered Series).

ACCEPTED MANUSCRIPT

Appendix 2. CCM Test Results



Noisy-t10-ccm | Noisy-t11-ccm | Noisy-t12-ccm | Noisy-t13-ccm

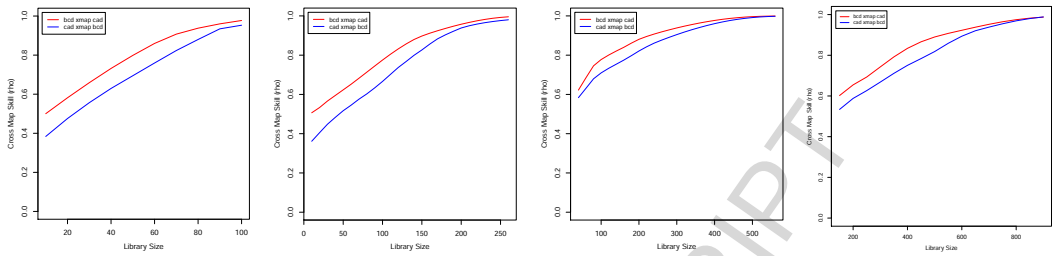
Noisy-t14(1)-ccm | Noisy-t14(2)-ccm | Noisy-t14(3)-ccm | Noisy-t14(4)-ccm

Noisy-t14(5)-ccm | Noisy-t14(6)-ccm | Noisy-t14(7)-ccm | Noisy-t14(8)-ccm

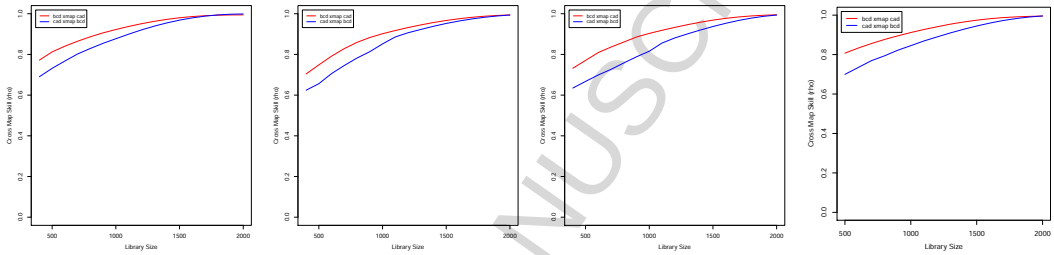
Figure 18: CCM test results for Bcd and Cad (Noisy Series).

ACCEPTED MANUSCRIPT

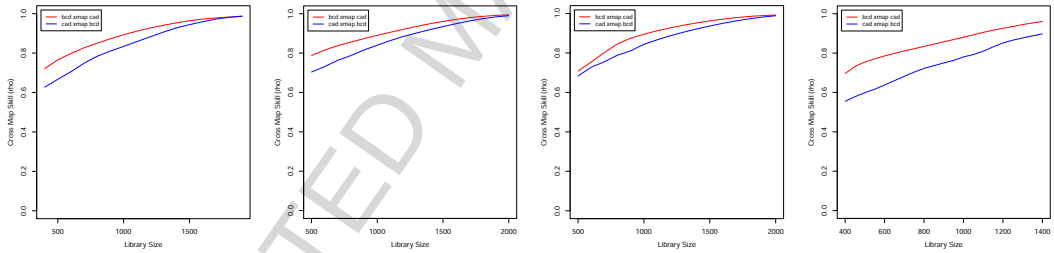
ACCEPTED MANUSCRIPT



Filtered-t10-ccm | Filtered-t11-ccm | Filtered-t12-ccm | Filtered-t13-ccm



Filtered-t14(1)-ccm | Filtered-t14(2)-ccm | Filtered-t14(3)-ccm | Filtered-t14(4)-ccm



Filtered-t14(5)-ccm | Filtered-t14(6)-ccm | Filtered-t14(7)-ccm | Filtered-t14(8)-ccm

Figure 19: CCM test results for Bcd and Cad (Filtered Series).

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

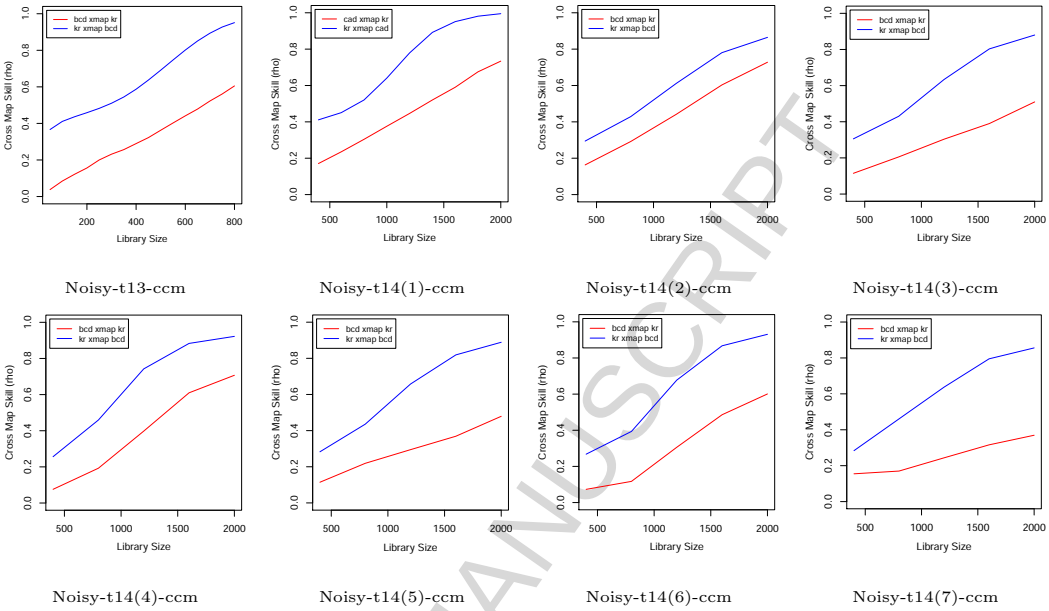


Figure 20: CCM test results for Bcd and Kr (Noisy Series).

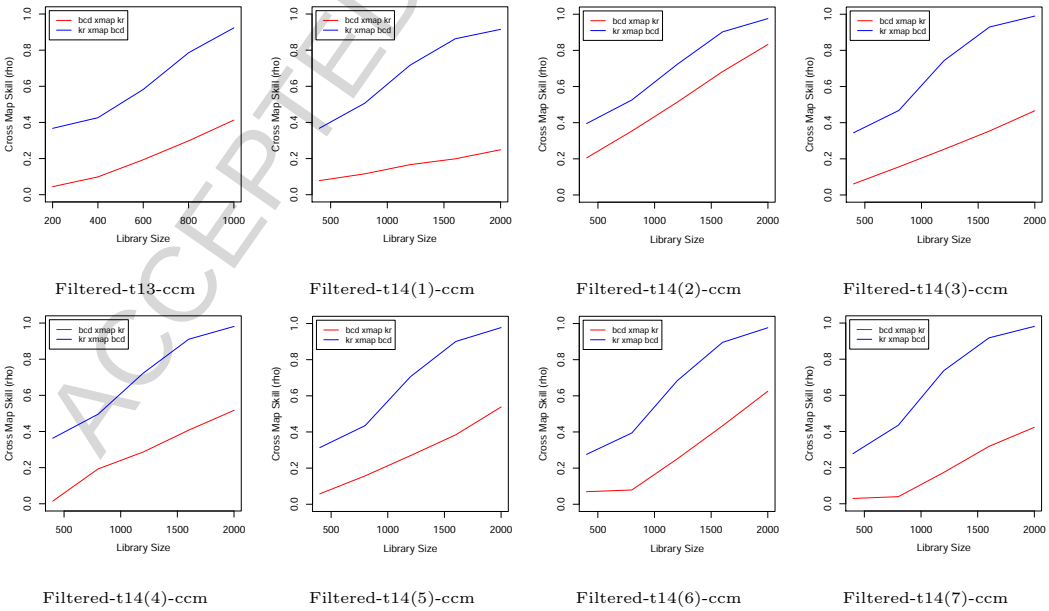


Figure 21: CCM test results for Bcd and Kr (Filtered Series).

ACCEPTED MANUSCRIPT

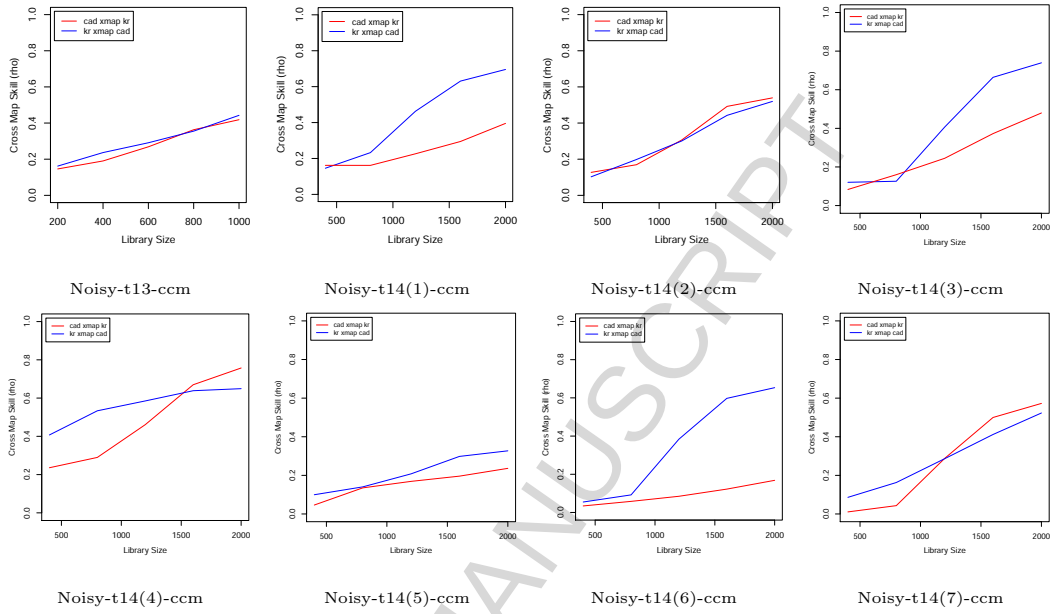


Figure 22: CCM test results for Cad and Kr (Noisy Series).

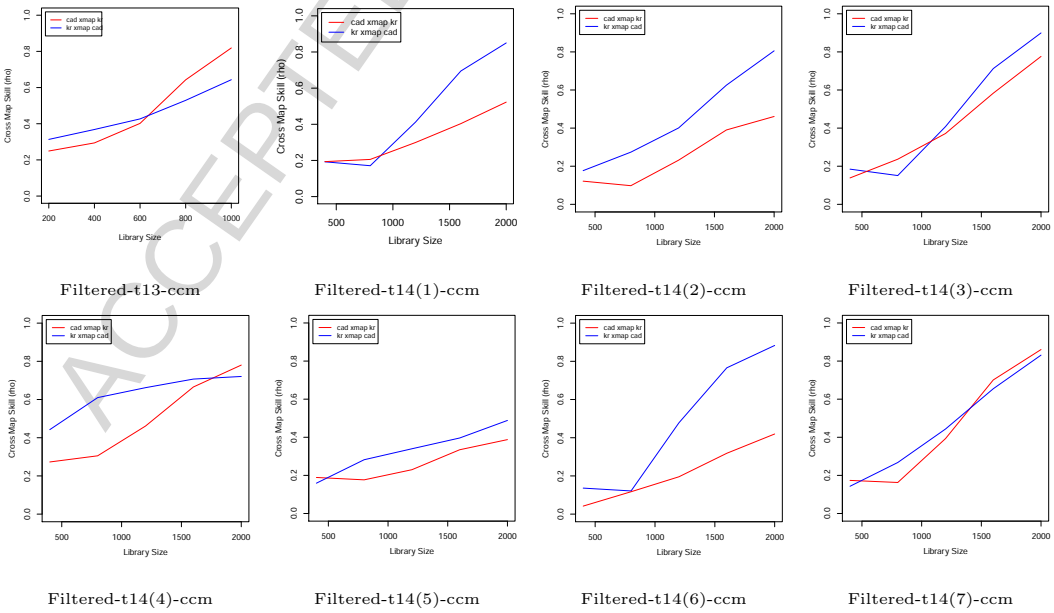


Figure 23: CCM test results for Cad and Kr (Filtered Series).

View publication stats